Sistemas biométricos multimodales que emplean rasgos audio-visuales

MSc. Susana C. Romaniz

Grupo de Investigación en Seguridad de las Tecnologías de Información y Comunicaciones
Facultad Regional Santa Fe - Universidad Tecnológica Nacional
sromaniz@frsf.utn.edu.ar

Resumen. Se observa en la actualidad una tendencia muy marcada al empleo de técnicas biométricas (las que incluyen rostro, huellas digitales, geometría de la mano, iris, patrones de retina, firma y voz, entre los más destacados) en los sistemas de identificación/autenticación de personas. Todas estas técnicas presentan diferentes grados de singularidad, permanencia, mensurabilidad, desempeño, aceptacción del usuario y robustez. Por otra parte, diferentes investigaciones sobre sistemas biométricos multimodales han demostrado que mediante su empleo es posible mejorar la carencias de cualquier sistema biométrico unimodal, por lo que desde hace algunos años, proliferan las propuestas para el empleo de dos o más métodos biométricos independientes. En el presente trabajo se analizan aspectos significativos de estos métodos y aproximaciones y se desarrollan breves reseñas, a partir de la búsqueda bibliográfica. Los aspectos abordados incluyen la fusión de información (poniendo el acento especialmente en la identificación y verificación audio-visual de personas), y las técnicas adaptativas y no-adaptativas para la decisión de verificación (utilizando información de habla y de rostro) en condiciones de audio con ruido.

Palabras Claves: biometría, biometría multimodal, seguridad de la información.

1. INTRODUCCIÓN

Se observa en la actualidad una tendencia muy marcada al empleo de técnicas biométricas (las que incluyen rostro, huellas digitales, geometría de la mano, iris, patrones de retina, firma y voz, entre los más destacados) en los sistemas de identificación/autenticación de personas. Todas estas técnicas presentan diferentes grados de singularidad, permanencia, mensurabilidad, desempeño, aceptacción del usuario y robustez [1].

Un sistema de verificación biométrica (o autenticación en la jerga de la seguridad de la información) verifica la identidad de un reclamante en base a atributos biométricos de una persona. Además de las diferentes formas de control de acceso (por ejemplo, control fronterizo, acceso a información), los sistemas de verificación también resultan de utilidad en trabajos forenses, en los que la tarea consiste en determinar si un muestra biométrica dada pertenece a un determinado sospechoso.

Por otra parte, diferentes investigaciones sobre sistemas biométricos multimodales han demostrado que mediante su empleo es posible mejorar la carencias de cualquier sistema biométrico unimodal, por lo que desde hace algunos años, proliferan las propuestas para el empleo de dos o más métodos biométricos independientes:

- combinando evidencia de la verificación de un hablante y el reconocimiento de su rostro [2],
- empleando un esquema de fusión a nivel abstracto denominado "*2-from-3-approach*", en el que se integran rostro, movimiento de labios y habla, que se basa en el principio de que las personas utilizan múltiples indicios para identificar a una persona [3],
- empleando una estrategia de integración que se focaliza en múltiples vistas instantáneas de una única propiedad biométrica utilizando un framework bayesiano [4],
- combinando datos biométricos (por ejemplo voz grabada) con datos no-biométricos (por ejemplo una contraseña) [5],
- integrando dentro de un sistema biométrico multimodal rostro, huella digital y habla para realizar una identificación personal [1].

En el presente trabajo se analizan aspectos significativos de estos métodos y aproximaciones y se des-

arrollan breves reseñas, a partir de la búsqueda bibliográfica. Los aspectos abordados incluyen la fusión de información (poniendo el acento especialmente en la identificación y verificación audio-visual de personas), y las técnicas adaptativas y no-adaptativas para la decisión de verificación (utilizando información de habla y de rostro) en condiciones de ruido de audio.

2. ASPECTOS GENERALES

Tradicionalmente, se han utilizado las contraseñas (seguridad basada en conocimiento) y las tarjetas de identificación (seguridad basada en token) para restringir el acceso a diferentes tipos de aplicaciones. Sin embargo, se puede fácilmente quebrar la seguridad en las mismas cuando se divulga una contraseña a un usuario no autorizado o un impostor roba una credencial. El surgimiento de la biometría en el campo de la identificación/autenticación de personas (seguridad basada en lo que se es) permite resolver los problemas que debilitan los métodos tradicionales de verificación.

La biometría hace referencia a la identificación (o verificación) automática de una persona (o una identidad reclamada) mediante el empleo de ciertos rasgos fisiológicos o de comportamiento asociados con la persona. Por esta razón, los sistemas biométricos presentan la ventaja de no poder ser fácilmente robados o compartidos respecto de los métodos tradicionales de seguridad.

Un sistemas de autenticación basado en biometría opera en dos modos:

1. **Modo Registración** (*enrollment*), en el que se adquieren los datos biométricos del usuario utilizando un lector biométrico y se almacenan los datos en una base de datos etiquetados con una identidad del usuario para facilitar la autenticación.
2. **Modo Autenticación** (*authentication*), en el que nuevamente se adquieren los datos biométricos del usuario y el sistema los utiliza para identificar quién es el usuario, o para verificar la identidad reclamada del usuario; la identificación comprende la comparación de la información biométrica adquirida contra plantillas correspondientes a todos los usuarios existentes en la base de datos, y la verificación comprende la comparación sólo con aquellos datos que corresponden a la identidad reclamada. En consecuencia, la identificación y la verificación son dos problemas diferentes que tienen sus propias complejidades.

Un **sistema biométrico simple** tiene cuatro componentes principales:

1. Módulo sensor (sensor module), encargado de la adquisión de los datos biométricos de un individuo.
2. Módulo de extracción de características (feature extraction module), en el cual se procesan los datos adquiridos para extraer los valores de la característica.
3. Módulo de correspondencia (matching module), en el que se comparan los valores de característica contra los almacenados para generar un puntaje de correspondencia.
4. Módulo de toma de decisión (decision-making module), encargado de establecer la identidad del usuario o de aceptar/rechazar una identidad reclamada en base al puntaje de correspondencia generado en el módulo de correspondencia.

El desempeño de un sistema bimétrico se puede medir mediante reportes FAR -*False Acceptance Rate*- y FRR -*False Rejection Rate*- en diferentes umbrales. Estos dos factores generalmente se representan en una curva ROC -*Receiver Operating Characteristic*-; alternativamente, se puede graficar la tasa de aceptación genuina con respecto a la FAR.

Tanto FAR como FRR se computan generando todos los puntajes de correspondencia genuinos y los de impostor, y luego se establece un umbral para decidir si aceptar o rechazar una correspondencia. Se obtiene un puntaje de correspondencia genuina cuando se comparan dos vectores de característica que corresponden al mismo individuo, y se obtiene un de correspondencia de impostor cuando se comparan vectores de característica que corresponden a dos individuos diferentes.

El desempeño de un sistema biométrico se ve fuertemente afectado por la confiabilidad del sensor que

se utilice y los grados de libertad que ofrecen las características extraídas de la señal sensada. Además, si el rasgo biométrico sensado o medido presenta ruido (por ejemplo, una huella digital con una cicatriz o un voz alterada por un resfrío), el puntaje de correspondencia resultante que calcule el módulo de correspondencia no será confiable. Dicho de manera simple, el puntaje de correspondencia generado por una entrada ruidosa posee una amplia variación; este problema se puede resolver mediante la instalación de múltiples sensores que capturen diferentes rasgos. Se espera que estos sistemas, conocidos como sistemas biométricos multimodales, resulten más confiables debido a la presencia de múltiples porciones de evidencia; asimismo, estos sistemas son capaces de satisfacer los requerimientos severos de desempeño impuestos por algunas aplicaciones.

Los sistemas biométricos multimodales resuelven el problema de la no-universalidad, ya que es posible que un subconjunto de usuarios no posea una biometría particular; por ejemplo, el módulo de extracción de características de huellas digitales puede ser incapaz de extraer características de huellas digitales asociadas con individuos específicos debido a la pobre calidad de los surcos; en tales circunstancias, resulta útil adquirir múltiples rasgos biométricos para la verificación de una identidad.

Además proveen medidas anti-falsificación, dificultándole las acciones al intruso, que debe falsificar simultáneamente múltiples rasgos biométricos. Y al solicitarle al usuario la presentación de un subconjunto aleatorio de rasgos biométricos, el sistema asegura que un usuario "vivo" está presente en el punto de adquisición.

Pero estos sistemas requieren de un esquema de integración para fusionar la información presentada por las modalidades particulares.

3. RESEÑA SOBRE LA FUSIÓN DE INFORMACIÓN

Hablando en términos amplios, **fusión de información** (*information fusion*) comprende cualquier área que se ocupa de la utilización de una combinación de diferentes fuentes de información, ya sea para generar una formato representacional o para tomar una decisión. Esto incluye: construcción de consenso, teoría de decisión en equipo, integración de múltiples sensores, fusión de datos multimodales, combinación de múltiples expertos/clasificadores, detección distribuida, y toma de decisiones distribuida. Los primeros trabajos sobre la materia aparecen a principios de los años 80. [6,7,8,9].

Cuando se lo analiza desde el punto de vista de la toma de decisiones, existen varios motivos por los que utilizar fusión de información:

- Utilización de información complementaria (por ejemplo, audio y video) pueden reducir las tasas de error.
- Empleo de múltiples sensores (es decir, redundancia) puede incrementar la confiabilidad.
- Costo de implementación reducido por el empleo de varios sensores más baratos que un único sensor de costoso.
- Sensores físicamente separados, permitiendo la adquisición de información desde diferentes puntos de vista.

Las personas emplean a diario la fusión de información; algunos ejemplos que se pueden mencionar son: el uso de ambos ojos, ver y escuchar el mismo objeto, o ver y escuchar a una persona hablar (lo cual mejora la inteligibilidad en ambientes ruidosos). Existen diferentes métodos para realizar la fusión de información, los que se suelen dividir en varias categorías: fusión a nivel de datos de sensor, fusión a nivel de características, fusión de puntaje, y fusión de decisión.

No obstante, resulta más intuitivo clasificarlos en tres categorías principales:

- **Fusión pre-mapeo**. La información se combina antes de cualquier empleo de expertos o clasificadores.
- **Fusión en medio del mapeo**. La información se combina durante el mapeo desde el espacio sensor-data/característica hacia el espacio opinión/decisión.

- **Fusión post-mapeo**. La información se combina luego del mapeo desde el espacio sensor-data/característica hacia el espacio opinión/decisión (en este caso el mapeo se realiza mediante la combinación de expertos o clasificadores en cada posible decisión).

En la fusión pre-mapeo, existen dos sub-categorías principales:

- Fusión a nivel de datos de sensor.
- Fusión a nivel de característica.

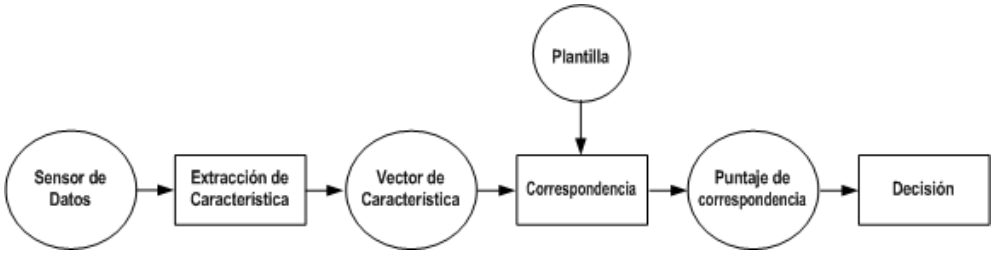
En la fusión post-mapeo, también existen dos sub-categorías principales:

- Fusión de decisión.
- Fusión de opinión, también se la denomina fusión de puntaje.

4. COMPARACIÓN DE MÉTODOS DE FUSIÓN MÁS DIFUNDIDOS EN SISTEMAS MULTIMODALES

Como se sugiere en la literatura (por ejemplo en [10,11]), los sistemas multimodales más difundidos que hace uso de biometrías múltiples se categorizan en tres arquitecturas de acuerdo a las estrategias utilizadas para la fusión de información (estas categorías son consistentes con las indicadas para los sistemas de fusión de información generales, que se describen en la sección anterior.):

- Fusión a nivel de extracción de características
- Fusión a nivel de puntaje de correspondencia
- Fusión a nivel de decisión



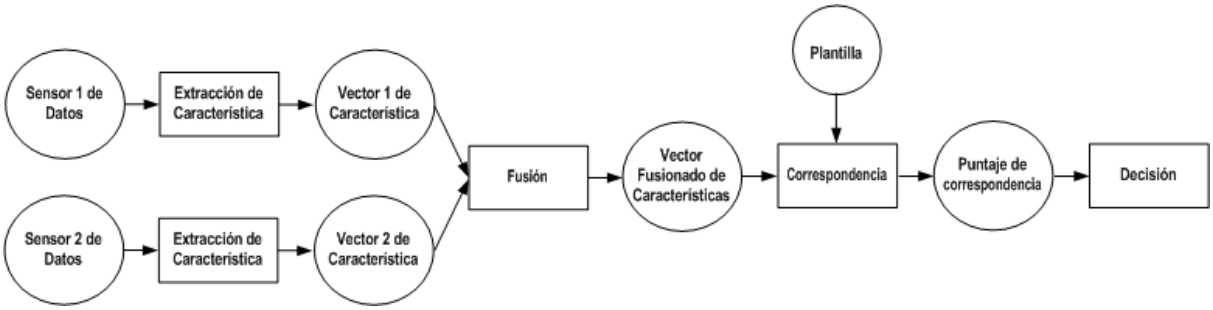
Los sistemas se clasifican de acuerdo a cuán temprano se combina la información proveniente de los diferentes sensores durante el proceso de autenticación. Las autenticación biométrica es un proceso en cadena [12], como se describe en la figura anterior. A continuación se analizan cada una de las tres arquitecturas y se analizan las actividades de investigación relacionadas con las mismas.

3.1 Fusión en el nivel de extracción de las características

En esta arquitectura, la información se extrae desde diferentes sensores, y se la codifica dentro de un vector de característica fusionado; luego, se los compara con plantilla almacenada (la que es asímismo un vector de característica fusionado que se encuentra almacenado en la base de datos) y se le asigna un puntaje de correspondencia, al igual que en un sistema biométrico unimodal.

Las búsquedas bibliográficas realizadas no revelan la existencia de investigaciones significativas reciente relativas a esta método de fusión, lo que sugiere que se lo prefiere menos que los otros dos métodos. Esto puede deberse a dos problemas que presenta:

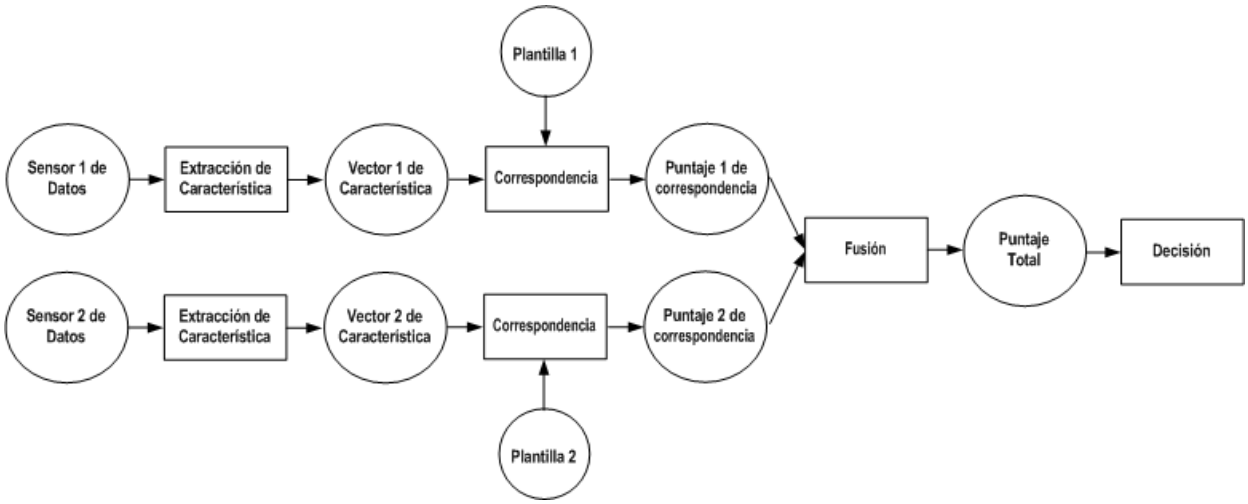
1. los vectores de característica que se deben fusionar pueden ser incompatibles (por ejemplo, debido a problemas numéricos) o algunos de ellos podrían no estar disponibles (por ejemplo, en casos donde el usuario no posee todos los identificadores biométricos); en tanto el primero de los problemas se puede solucionar con un diseño más complejo del sistema, lo que conduce a un sistema muy fuertemente acoplado, el segundo provoca problemas en la registración que ya existen en los sistemas biométricos unimodales.



2. la generación del puntaje es problemática, ya que aún en el caso de un sistema biométrico unimodal, resulta demasidado dificultoso encontrar un buen clasificador, es decir, generar un puntaje representativo basado en la correspondencia de un vector de característica y los datos de una plantilla; cuando se trata de vectores de característica fusionados de grandes dimensiones, esto es aún más complicado, ya que la relación entre los diferentes componentes de dicho vector fusionado puede no ser linear [13].

3.2 Fusión en el nivel de puntaje de correspondencia

En un sistema biométrico multimodal que se construye con esta arquitectura, los vectores de característica se crean independientemente para cada sensor, y luego se comparan con las plantillas almacenadas en forma separada para cada uno de los rasgos biométricos. En base a la proximidad del vector de carac terística y la plantilla, cada subsistema calcula su propio puntaje de correspondencia. Finalmente, estos valores individuales se combinan en un puntaje total que se pasa al módulo de decisión.



El flujo de proceso dentro de un subsistema es el mismo que en un sistema biométrico unimodal, lo que permite el empleo de algoritmos ya probados para la extracción de características y la determinación de correspondencia.
Se destacan dos informes de investigación, [11,14], en los que se incorporan en un único sistema de autenticación método de exploración de rostro, verificación de huellas digitales y exploracion de geometría de mano; en los mismos se emplean métodos bien conocidos para cada identificador; luego, se normalizan y combinan los puntajes de correspondencia para las tres modalidades utilizando alguno de los siguientes métodos:

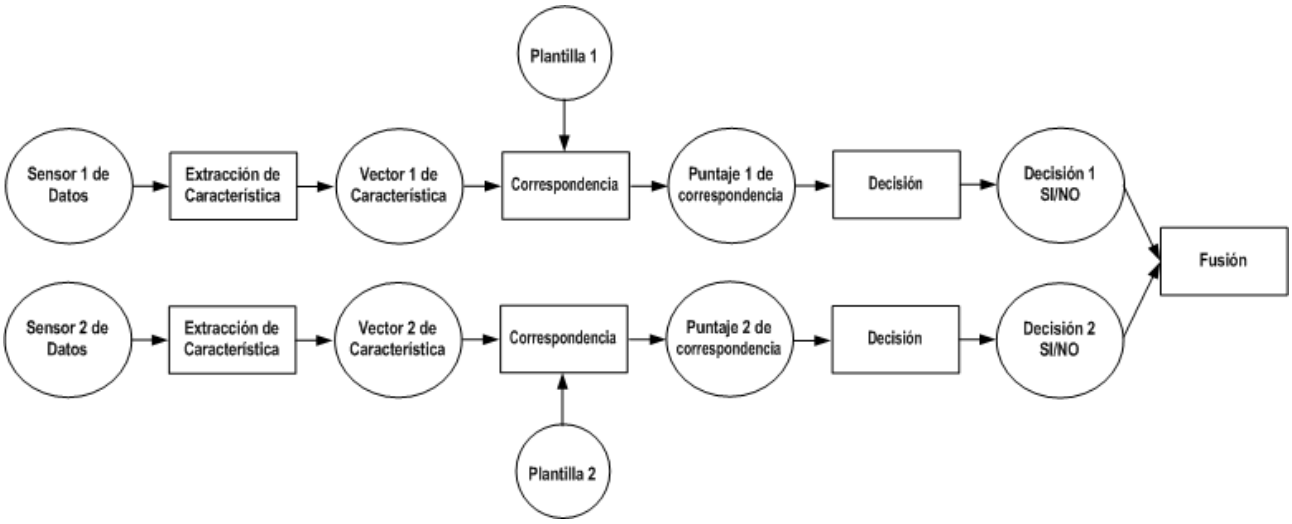
- Suma ponderada, calcula el promedio ponderado de los puntajes.

- Árbol de decisión, emplea para los diferentes puntajes una secuencia de comparación de umbrales para tomar una decisión de autenticación.
- Análisis de discriminante linear, transforma los vectores de 3-dimensiones de puntajes en un nuevo sub-espacio, en el que está maximizada la separación entre los puntajes de las clases reclamante verdadero e impostor; los parámetros óptimos para esta transformación se calculan en forma anticipada en base a un conjunto de datos de entrenamiento.

En base a los resultados experimentales, las primeras conclusiones indican que el método de suma ponderada logra el mejor desempeño.
Además, se suman al sistema reglas de aprendizaje: inicialmente, se asignan las mismas ponderaciones a cada rasgo biométrico, los cuales se modifican luego cada vez que se utilizan, a los fines de minimizar las tasas de falsos positivos y falsos negativos.
Si bien la novedad de la estrategia que hace uso de ponderaciones específicas del usuario, resulta prometedora su aplicación para hacer frente a problemas de rasgos biométricos no-universales y de la plantilla; si un usuario no posee cierto identificador biométrico y sólo posee características débiles, es posible ajustar la ponderación para reducir su influencia.

3.3 Fusión en el nivel de decisión

En esta arquitectura, se toma una decisión de autenticación separada para cada rasgo biométrico; luego, estas decisiones se combinan en un voto final.



La fusión en el nivel de decisión resulta en una arquitectura de sistema débilmente acoplado, en la que cada subsistema se ejecuta como un sistema biométrico unimodal, lo que hace que resulte muy atractivo para los fabricantes que muchas veces lo presentan bajo la denominación de "biometría en capas", concepto que se encuentra respaldado por la aparición de estándares biométricos tales como BioAPI [15].
Existen variadas estrategias para combinar diferentes decisiones en una decisión de autenticación final, las que van desde mayoría de votos hasta métodos estadísticos más sofisticados [13].
Tomando como ejemplo a *BioNetrix Authentication Suite*, se tiene la siguiente combinación de estrategias (en [16] se incluye una lista muy completa de posibles combinaciones alternativas):

- Operador AND, requiere de una decisión positiva de todos los módulos de verificación;
- Operador OR, intenta autenticar al usuario utilizando un rasgo biométrico; si falla, ofrece otro intento con otro módulos de verificación;
- Operador RANDOM, selecciona aleatoriamente un rasgo biométrico; si bien se trata de una idea muy simplista, hace mucho más difícil engañar al sistema.

La fusión en el nivel de decisión es una etapa muy tardía del proceso de autenticación, por lo que se presume que no presenta el mismo potencial de mejora del desempeño del sistema global como la fusión en el nivel de puntaje de correspondencia.

5. SISTEMAS QUE EMPLEAN RASGOS AUDIO-VISUALES

A continuación se reseñan brevemente las principales contribuciones realizadas en este campo, tanto en lo que hace a la identificación como la verificación de identidades de personas. Se distinguen dos categorías principales de métodos: no-adaptativos y adaptativos. En el primero de los métodos, la contribución de cada experto se establece a priori, mientras que en el segundo, la contribución de al menos un experto varía de acuerdo a su confiabilidad y capacidad de discriminación en presencia de alguna condición ambiental (por ejemplo, la contribución de un experto en habla se decrementa cuando baja la SNR -*Signal Noise Ratio*- del audio).

5.1 Métodos no-adaptativos

La fusión de información de audio y visual se ha aplicado al reconocimiento automatizado de personas desde las primeras propuestas de sistemas multimodales [17,18,2].

En [17], se combina información de imágenes de rostros y grabaciones de habla empleando fusión de suma ponderada:

$$f = w_1 o_1 + w_2 o_2$$

donde o_1 y o_2 son las opiniones de los expertos de rostro y de habla, respectivamente, con sus correspondientes ponderaciones, w_1 y w_2. Cada opinión refleja la probabilidad de que un reclamante sea el reclamante verdadero (es decir que una opinión baja sugiere que el reclamante es un impostor, en tanto que una opinión alta sugiere que el reclamente es el reclamente verdadero). Debido a la restricción sobre las ponderaciones, $\sum_{i=1}^{2} w_i = 1$, la ecuación anterior se reduce a:

$$f = w_1 o_1 + (1 - w_1) o_2$$

La verificación de la decisión se logra estableciendo umbrales de la opinión fusionada. Los resultados obtenidos de EER -*Equal Error Rate*- al emplear un único experto (habla 3.4%, rostro 3.0%) son significativamente superiores a los que se obtienen con el empleo de ponderaciones óptimas y umbrales (1,5%).

En [18] se combinan las opiniones de un experto de rostro (el que hace uso de características obtenidas a partir de imágenes estáticas frontales) y de un experto de habla, y se emplea el método de producto ponderado:

$$f = (o_1)^{w_1} \times (o_2)^{(1-w_1)}$$

Cuando el experto de habla se utiliza solo (es decir, w_1=1), se obtiene una tasa de identificación del 51%, mientras que cuando se emplea el experto de rostro solo (es decir, w_1=0), se obtiene una tasa de identificación del 92%; y utilizando una ponderación óptima, la tasa de identificación llega al 95%.

En [2] se emplean para la identificación de personas dos expertos de habla (para características estáticas y delta) y tres expertos en rostro (para las área de ojos, nariz y boca), utilizando el método de producto ponderado para la fusión de opiniones, donde las ponderaciones se determinaban en base a una heurística. Con los expertos estático y dinámico, se obtienen tasas de identificación del 77% y 71%, respectivamente; combinando los dos expertos de habla, este valor se incrementa al 88%. Con los expertos de rostro, se obtienen tasas de identificación del 80%, 77% y 83%, respectivamente; combinándolos, la tasa se incrementa al 91%. Cuando se combinan los cinco expertos, la tasa de identificación se incrementa al 98%.

En [3] se emplean tres expertos (de rostro frontal, de imagen dinámica de labios y de habla dependiente del texto), con un esquema de fusión híbrida en el que intervienen mayoría de votos y fusión de opinión; dos de los expertos deben acordar respecto de la decisión, y la opinión combinada tiene que exceder un umbral preestablecido. Este esquema presenta un mejor desempeño que cuando se utilizan dichos expertos en forma individual.

En [19] se emplea un experto de rostro frontal, que proporciona una opinión para cada una de las imágenes; cuando se utilizan múltiples imágenes de una persona para generar múltiples opiniones, éstas se fusionan mediante diferentes esquemas (entre los que se incluyen un caso especial de fusión por suma ponderada). De demuestra una reducción en las tasas de errores del 40%, y que las ganancias en el desempeño se tienden a saturar luego de utilizar cinco imágenes; estos resultados sugieren que el uso de una secuencia de video del rostro, en lugar de una imagen, provee un desempeño superior.
En [20] se intenta proporcionar fundamentos teóricos a los métodos más comunes de fusión, tales como métodos de suma y producto; sin embargo, los autores admiten que los supuestos utilizados "no son realistas para la mayoría de las aplicaciones". Los resultados experimentales para la combinación de tres expertos (dos de rostro -frontal y perfil- y uno de habla dependiente del texto) demuestran que el método de suma supera al de producto. En [21] se investiga la combinación de información de audio (habla) y visual (labios) mediante concatenación de vector de característica. A fin de hacer corresponder las tasas de tramas de ambas características, se extrae la información de habla a una tasa de 30 fps en lugar de los 100 fps tradicionales. En la configuracion dependiente del texto, el proceso de fusión presenta una mejora menor en el desempeño; sin embargo, en la configuración independiente del texto, el desempeño disminuye ligeramente, y se sugiere que el método concatenación de vector de característica es poco fiable.
En [22,23] se emplea una forma de fusión de suma ponderada para combinar dos expertos de opiniones: un experto en habla dependiente del texto y un experto en labios dependiente del texto. Utilizando una ponderación óptima, la fusión conduce a un mejor desempeño frente respecto del uso de dichos expertos en forma independiente.
En [24] se utiliza un experto de huellas digitales y un experto de rostro frontal, y se emplea un esquema de fusión híbrida que comprende fusión de lista ordenada y fusión de opinión: las opiniones del experto de rostro correspondientes a n identidades se combinan con las opiniones del experto de huella digital para las identidades correspondientes utilizando una forma del método de producto. Se utiliza este método híbrido a los fines de tener en cuenta la relativa complejidad computacional del experto de huellas digitales (significativamente más lento). Se demuestra que, en todos los casos testeados, la fusión presenta un mejor desempeño que cuando se emplean cualquier de los expertos solos.
En [25] se propone el uso de un post-clasificador bayesiano para alcanzar la decisión de verificación; formalmente, la regla se decisión se expresa como:

$$class = \begin{cases} C_1 if \prod_{i=1}^{N_E} p(o_i|\lambda_{i,true}) > \prod_{i=1}^{N_E} p(o_i|\lambda_{i,imp}) \\ C_2 otherwise \end{cases}$$

donde C_1 y C_2 son las clases reclamante verdadero e impostor, respectivamente, N_E es el número de expertos, en tanto que $\lambda_{i,true}$ y $\lambda_{i,imp}$ son, para el i-ésimo experto, los modelos paramétricos de la distribución de opiniones para el reclamante verdadero y el impostor, respectivamente. Debido a problemas de precisión en una implementación computacional, resulta más conveniente el empleo de una suma en lugar de series de multiplicaciones, y dado que la función logarítmica es una función monótona creciente, se puede modificar la regla de decisión de la siguiente manera:

$$class = \begin{cases} C_1 if \sum_{i=1}^{N_E} \log p(o_i|\lambda_{i,true}) - \sum_{i=1}^{N_E} \log p(o_i|\lambda_{i,imp}) \\ C_2 otherwise \end{cases}$$

La regla de decisión anterior, en la práctica, se modifica introduciendo un umbral a fin de permitir el ajuste de FAR y de FRR:

$$class = \begin{cases} C_1 if \sum_{i=1}^{N_E} \log p(o_i|\lambda_{i,true}) - \sum_{i=1}^{N_E} \log p(o_i|\lambda_{i,imp}) > t \\ C_2 otherwise \end{cases}$$

Además, se utilizan tres expertos, observándose que el uso del clasificador anterior (con distribuciones Beta) proporciona menores tasas de error que cuando se utilizan los expertos solos.
Los clasificadores que se investigaron son: SVM (*Support Vector Machine*), clasificador bayasiano (uti-

lizando distribuciones Beta), Discriminante Linear de Fisher, Árbol de Decisión y *Percepton* Multicapa; en cuanto a los expertos, se emplearon tres: un experto de rostro frontal, y dos expertos de habla (dependiente e independiente del texto). Se determina que el clasificador SVM y el bayasiano presenta los mejores resultados.

En [26] también investigan, para la fusión de opinión, varios clasificadores binarios y los métodos de fusión mayoría de votos y operadores AND y OR (lo que lleva a la categoría de fusión de decisión). Se utilizan tres expertos: experto de rostro frontal, experto de rostro de perfil y experto de habla independiente del texto. En el caso de fusión de decisión, cada experto actua como un clasificar, que provee una decisión "dura" en lugar de una opinión. Los clasificadores que se investigan son: Árbol de Decisión, *Percepton* Multicapa, clasificador basado en *Logistic Regression*, clasificador bayasiano utilizando distribuciones gaussianas, Discriminante Linear de Fisher, y varias formas del clasificador *k-Nearest Neighbour*. Se determinó que el clasificador basado en *Logistic Regression* proporciona la tasa de errores más baja y que resulta el más fácil de entrenar.

En [27] se utiliza el método suma ponderada para combinar las opiniones de un experto de habla y un experto de labios (ambos independientes del texto); el desempeño del primero se disminuye deliberadamente variando las cantidad de ruido blanco en los datos de habla. Los resultados experimentales demuestran que si bien el desempeño del sistema siempre es mejor que cuando se emplea sólo el experto de habla, el mismo disminuye a medida que se incrementa el nivel de ruido. De acuerdo a los valores de ponderación (que se seleccionan previamente), el desempeño con altos niveles de ruido son realmente peores que cuando se utiliza el experto de labios solo.

Se propone un método basado en estadísticas para la selección de las ponderaciones, que da por resultado un buen desempeño bajo condiciones limpias, y nunca cae por debajo del desempeño de un experto de labios en condiciones ruidosas; sin embargo, el desempeño bajo condiciones ruidosas no fue óptimo.

La ponderación para el experto de habla se calcula de la siguiente manera:

$$w_1 = \frac{\varsigma_2}{\varsigma_1 + \varsigma_2} \quad \text{donde} \quad \varsigma_i = \sqrt{\frac{\sigma^2_{i,true}}{N_{true}} + \frac{\sigma^2_{i,imp}}{N_{imp}}}$$

siendo, para el *i*-ésimo experto, ζ_i el error estándar de la diferencia entre las medias $\mu_{i,true}$ y $\mu_{i,imp}$ de las opiniones para el reclamos verdadero e impostor, $\sigma^2_{i,true}$ y $\sigma^2_{i,imp}$ las correspodientes varianzas, y N_{true} y N_{imp} el número de opiniones para los reclamos verdadero e impostor, respectivamente.

Se asume que el error estandar representa la indicación relativa de la capacidad de discriminación de un experto; cuanto menor variación exista en las opiniones para reclamos conocidos, menor será el error estandar; y, en consecuencia, un error estandar bajo indica un mejor desempeño.

En [28] se evalúan variaciones del *Multi-Stream Hidden Markov Models* -MS-HMMs-, una forma de fusión en medio del mapeo, en la tarea de identificación de una persona por medios audio-visuales dependientes del texto. El flujo de audio consta de una secuencia de vectores que contienen *Mel Frecuency Cepstral Coefficients* -MFCCs- [29] y sus deltas [30], en tanto que el flujo de video consta de una secuencia de vectores de característica que describen el contorno de los labios. Debido a la naturaleza de la implementación MS-HMM, la tasa de tramas de características de video debe concordar con la tasa de tramas de caracterísiticas de audio. Se realizan pruebas utilizando una pequeña base de datos audio-visuales, las que demuestran que para altos SRNs, el desempeño es comparable con un sistema que sólo emplea HMM de audio, mientras que con bajos SRNs, el sistema multi-flujo presenta un desempeño significativamente superior al sistema que sólo utiliza audio y excede al desempeño del sistema que sólo empleo video. Este trabajo no incluye una comparación con los sistemas que emplean fusión pre-mapeo o post-mapeo, por ejemplo, utilizando dos expertos diferentes y fusión de opinión.

En [31] se resuelven varias limitaciones existentes en los sistemas MS-HMM previos, permitiendo que los dos flujos se encuentren desincronizados en el tiempo (debido a que los eventos relacionados con los flujos pueden comenzar y/o finalizar en puntos diferentes) y que presenten diferentes tasas de tramas. Se

realizan pruebas sobre una pequeña base de datos audio-visual, y se emplean dos flujos de características similares a los descriptos en [28]; se observa que para SNRs relativamente altos, el desempeño es peor que cuando se emplea un sistema de audio dependiente del texto, mientras que para SNRs menores se mejora el desempeño (y el sistema resulta más robusto) que un sistema HMM dependiente de texto que emplea concatenación de vector de característica.

5.2 Métodos adaptativos

En [32] se extiende el trabajo presentado en [27] al proponer un método heurístico para ajustar las ponderaciones; los resultados experimentales muestran que, si bien decrese significativamente el desempeño a medida que se incrementa el nivel de ruido, siempre resulta mejor que utilizar solamente el experto de habla; sin embargo, se observa que con niveles altos de ruido, el empleo de ponderaciones iguales (no-adaptativo) ofrece un mejor desempeño. Una desventaja importante del método es que el cálculo de las ponderaciones demanda encontrar la opinión del experto de habla para todos los reclamos posibles (es decir, todas las personas registradas en el sistema), limitando de esta manera la solución a sistemas que poseen un número reducido de clientes debido a consideraciones prácticas (es decir, el tiempo que demanda verificar un reclamo). Es más, según se describe en [27], se observan limitaciones similares en ambientes experimentales.

En [33], los autores proponen otra técnica heurística para el ajuste de las ponderaciones; en una configuración dependiente del texto, el sistema presenta un desempeño siempre superior al que se tiene utilizando solamente el experto de labios; sin embargo, en una configuración independiente del texto, bajo condiciones de SNR bajo, el desempeño fue peor que cuando se utiliza sólo el experto de labios.

La ponderación para el experto de habla se calcula de la siguiente manera: $w_1 = \left[\frac{\varsigma_2}{\varsigma_1 + \varsigma_2}\right]\left[\frac{\kappa_1}{\kappa_1 + \kappa_2}\right]$ donde $\frac{\varsigma_2}{\varsigma_1 + \varsigma_2}$ se calcula según la ecuación ya indicada durante la etapa de entrenamiento y $\kappa_i = \frac{\left|M(o_i)_{i,true} - M(o_i)_{i,imp}\right|}{\mu_{i,true}}$ se calcula durante el testeo; para el experto *i*-ésimo, $M(o_i)_{i,true} = \frac{(o_i - \mu_{i,true})^2}{\sigma^2_{i,true}}$ es la distancia unidimensional cuadrática Mahalanobis entre o_i y el modelo de opiniones para los reclamantes verdaderos; además, $\mu_{i,true}$ y $\sigma^2_{i,true}$ son, respectivamente, la media y la varianza de las oponiones para reclamantes verdaderos, los que se determinan durante la fase de entrenamiento.

De manera similar $M(o_i)_{i,imp} = \frac{(o_i - \mu_{i,imp})^2}{\sigma^2_{i,imp}}$ es la distancia unidimensional cuadrática Mahalanobis entre la opinión o_i y el modelo de opiniones de los impostores; acá, $\mu_{i,imp}$ y $\sigma^2_{i,imp}$ son la media y la varianza de las opiniones para impostor, respectivamente, se los determina durante la etapa de entrenamiento.

Bajo condiciones limpias, la distancia entre una opinión dada para un reclamante verdadero y el modelo de opiniones correspondiente debe ser pequeña; de manera similar, la distancia para un reclamante verdadero y el modelo de opiniones para los impostor debería ser grande. Lo inverso se aplica a una opinión dada para un impostor; por ello, bajo condiciones limpias, κ_i debe ser grande. Se emplea una evidencia empírica para argumentar que bajo condiciones ruidosas, las distancias deben disminuir y por ello κ_i debe también disminuir.

En [34] se propone el siguiente método de ajuste de la ponderación; cada vez que se graba habla, generalmente la declaración está precedida por un breve segmento que sólo contiene ruido ambiental; a partir de cada declaración de entrenamiento, se utilizan los MFCCs [35,36] obtenidos del segmento de ruido para construir un GMM de ruido global, λ_{noise}; dado un testeo de habla grabada, se emplean los vectores de característica MFFC N_{noise} $\{\overline{x_i}\}_{i=1}^{N_{noise}}$, representando al segmento de ruido, para estimar la calidad de la declaración mediante la medición del desajuste respecto de λ_{noise} de la siguiente manera

$$q = \frac{1}{N_{noise}} \sum_{i=1}^{N_{noise}} \log p(\bar{x}_i \mid \lambda_{noise})$$

Cuanto mayor sea la diferencia entre las condiciones de entrenamiento y de testeo, menor ha de resultar q; entonces, q se mapea con un valor comprendido en el intervalo [0,1] utilizando una curva sigmoidal, donde a y b describen la forma de la curva:

$$q_{map} = \frac{1}{1 + \exp[-a(q-b)]}$$

estos valores se seleccionan en forma manual de tal manera que q_{map} sea próximo a 1 para declaraciones de entrenamiento limpias, y próximo a 0 para declaraciones de entranamiento corrompidas artificialmente con ruido.

Si se asume que el experto de rostro es el primer experto y que el de habla, el segundo, dada una ponderación previa $w_{2,prior}$ para el experto de habla (que se determina sobre datos limpios), la ponderación adaptada para el experto de habla se calcula de la siguiente manera: $w_2 = q_{map}\, w_{2,prior}$. Dado que se está utilizando un sistema de dos modalidades, la ponderación correspondiente para el experto de rostro se encuentra utilizando $w_1 = 1 - w_2$. Este método de ajuste de ponderaciones se denomina *detección de desajuste*.

6. CONCLUSIONES

En la actualidad, existe un fuerte consenso entre los investigadores y la industria que la tecnología multimodal será la piedra angular en el empleo masivo de la biometría en los campos de la identificación/verificación de personas. En este trabajo se han reseñado diferentes métodos de abordaje de sistemas biométricos multimodales, entre los que se destacan interesantes intentos por atemperar algunos de los problemas que aún hoy no se ha sido posible eliminar en los sistemas biométricos tradicionales; de estos intentos, los más promisorios aparentan ser los que utilizan fusión de información en el nivel de puntaje de correspondencia y que, además, incluyen ponderaciones asociadas a usuarios (o grupos) particulares así como umbrales tales como los propuestos en [11].

Y como ya se planteara anteriormente, resulta evidente que la adquisición de múltiples identificadores biométricos dificulta significativamente las acciones que debe realizar un impostor para engañar al sistema de identificación/verificación, ya que debe presentar múltiples muestras coordinadas creadas artificialmente.

Sin embargo, todos estos beneficios no se logran sin algún tipo de cargo, ya que estos sistemas son menos costosos, y presentan efectos significativos sobre sus usuarios, pudiendo resultando en una baja aceptación, en particular en lo que hace a cuestiones de privacidad y al inconveniente derivado de la adquisición multinivel de datos.

Muchas de las arquitecturas más prometedoras hoy aún se encuentran en un estadio experimental. Y las tecnologías ya disponibles poseen arquitecturas multicapas, con un acoplamiento débil entre los diferentes subsistemas, a tal punto que algunos casos presentan diferentes interfaces de usuario.

Por ello, hoy día la industria y, muy particularmente aquellos actores que aguardan que esta tecnología aportes significativos que impulsen la masividad de sistemas de información de seguridad crítica (gobiernos, salud, bancos, etc.), demandan de los investigadores y fabricantes la aparición de soluciones verdaderamente integradas y altamente confiables y, que al mismo tiempo, mejoren la facilidad de uso (más allá del empleo de múltiples identificadores biométricos).

REFERENCIAS

1. Jain, A., Bolle, R. and Pankanti, S. “Biometricas: Personal identification in networked society” 2Ed. Kluwer Academic Publishers. 1999.
2. Brunelli, R. and Falavigna, D. “Personal identification using multiple cues” IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 17, No. 10. 1995.

3. Dieckmann, U., Plankensteiner, P., and Wagner, T. "SESAM: A biometric person identification system using sensor fusion". Pattern Recognition Letters, Vol. 18, No. 9. 1997.
4. Kittler, J., Li, Y., Matas, J. and Sanchez, M. U. "Combining evidence in multi-modal personal identity recognition systems" Proceedings 1st Int. Conf. On Audio Video-Based Personal Authentication. Crans-Montana. 1997.
5. Maes S. and Beigi, H. "Open sesame! Speech, password or key to secure your door?". Proceedings 3rd Asian Conference on Computer Vision. Hong Kong. 1998.
6. Barniv, H., Casasent, D. "Multisensor image registration: Experimental verificación". Proceedings of the SPIE 292. 1981.
7. Pau, L.F. "Fusion of multisensor data in pattern recognition". Kittler, K., Fu, K.S. and Pau, L.F. Pattern Recognition Theory and Applications (Proceedings of NATO Advanced Study Institute), D. Reidel Publ. Holland. 1982.
8. Tenney, R.R., Sandell Jr., N.R. "Detection with distributed sensors". IEEE Trans. Aerospace and Electronic Systems 17. 1981.
9. Tenney, R.R., Sandell Jr., N.R. "Strategies for distributed decision making". IEEE Trans. on Systems, Man and Cybernetics 11. 1981.
10. Hong, L. et al. "Can Multibiometrics Improve Performance?". Proceedings AutoID. 1999.
11. Ross, A. and Jain, A. K. "Information Fusion in Biometrics". Pattern Recognition Letters. 2003.
12. Nanavati, Samir et al. "Biometrics: Identity Verification in a Networked World". Wiley Computer Publishing. New York. 2002
13. Prabhakar, S. and Jain, A. K. "Decision-level Fusion in Biometric Verification". Pattern Recognition v35 n4. 2002.
14. Jain, A. K. and Ross, A. "Learning User-specific Parameters in a Multibiometric System". Proceedings International Conference on Image Processing (ICIP). 2002.
15. Tilton, Catherine J. "An Emerging Biometric API Industry Standard". IEEE Computer v33 n2. 2000.
16. Speir, Michelle. "BioNetrix delivers layered biometrics suite". Federal Computer Week. 2000.
17. Chibelushi, C., Deravi, F. and Mason, J. "Voice and Facial Image Integration for Speaker Recognition". IEEE International Symposium and Multimedia Technologies and Future Applications. Southampton, UK. 1993.
18. Brunelli, R., Falavigna, D., Poggio, T. and Stringa, L. "Automatic Person Recognition Using Acoustic and Geometric Features". Machine Vision & Applications, Vol. 8. 1995.
19. Hall, D. and Llinas, J. "Multisensor data fusion". D. L. Hall and J. Llinas (Eds.), Handbook of Multisensor Data Fusion, CRC Press. USA. 2001.
20. Ho, T., Hull, J. and Srihari, S. "Decision combination in multiple classifier systems". IEEE Trans. Pattern Analysis and Machine Intelligence 16. 1994.
21. Hong, L. and Jain, A. "Integrating Faces and Fingerprints for Personal Identification". IEEE Trans. Pattern Analysis and Machine Intelligence 20. 1998.
22. Haigh, J. and Mason, J. "A voice activity detector based on cepstral analysis". Proceedings European Conf. Speech Communication and Technology. 1993.
23. Haigh, A. "Voice Activity Detection for Conversational Analysis". Masters Thesis, University of Wales. 1994.
24. Furui, S. "Recent advances in speaker recognition" Pattern Recognition Letters 18. 1997.
25. Abdeljaoued, Y. "Fusion of person authentication probabilities by Bayesian statistics". Proceedings 2nd Int. Conf. Audio- and Video-based Biometric Person Authentication. Washington D.C. 1999.
26. P. Verlinde, P. "A contribution to multi-modal identity verification using decision fusion". PhD Thesis, Department of Signal and Image Processing, Telecom Paris. France. 1999.
27. Wark, T., Sridharan, S. and Chandran, V. "Robust speaker verification via fusion of speech and lip modalities" Proceedings International Conf. Acoustics, Speech and Signal Processing. Phoenix. 1999.
28. Wark, T., Sridharan, S. and Chandran, V. "The use of temporal speech and lip information for multi-modal speaker identification via multistream HMM's". Proceedings International Conf. Acoustics, Speech and Signal Processing. Etambul. 2000.
29. Reynolds, D. "Experimental evaluation of features for robust speaker identification". IEEE Trans. Speech and Audio Processing 2. 1994.
30. Soong, F. and Rosenberg, A. "On the use of instantaneous and transitional spectral information in speaker recognition". IEEE Trans. Acoustics, Speech and Signal Processing 36. 1988.
31. Bengio, S. "Multimodal authentication using asynchronous HMMs". Proceedings 4th International Conf. Audio- and Video-based Biometric Person Authentication. Guildford. 2003.
32. Wark, T., Sridharan, S. and Chandran, V. "Robust speaker verification via asynchronous fusion of speech and lip information". Proceedins 2nd International Conf. Audio- and Video-based Biometric Person Authentication. Washington, D.C. 1999.
33. Wark, T. "Multi-modal speech processing for automatic speaker recognition". PhD Thesis, School of Electrical & Electronic Systems Engineering, Queensland University of Technology. Brisbane. 2000.
34. Sanderson, C. and Paliwal, K. "Noise compensation in a person verification system using face and multiple speech features". Pattern Recognition 36 (2). 2003.
35. Picone, J. "Signal modeling techniques in speech recognition". Proceedings of the IEEE 81. 1993.
36. Reynolds, D. "Experimental evaluation of features for robust speaker identification". IEEE Trans. Speech and Audio Processing 2. 1994.